Chapter 3

Participant Learning Outcomes

The statute mandates states to develop indicators for adults in:

- *Achievement in reading, writing, English language acquisition, problem-solving, and numeracy;*
- *Entry into postsecondary education, job retraining program, employment or career advancement, including the military; and*
- *Receipt of a secondary school diploma or GED.*

And for children in:

- *Improvement in ability to read on grade level or reading readiness;*
- *School attendance; and*
- *Grade retention and promotion.*

Chapter 3 focuses on participant outcomes—the achievements expected of all adults and children in any Even Start project. It describes how to measure and analyze outcomes in terms of the state-required indicators and how to develop additional project-identified outcomes. This chapter covers:

- state-required indicators;
- choosing additional outcomes at the local level;
- selecting instruments; and
- using outcome data for continuous improvement.

State-Required Indicators

The Even Start law broadly identifies the important learning outcomes expected of adults and children enrolled in Even Start projects. Although Congress defined the parameters, it requires each state to develop its own performance indicators.[1] Therefore, indicators vary by state. State-developed indicators identify educational achievements expected of adults and children enrolled in local projects. States report on these indicators every year in their consolidated reports, and some of these data can be aggregated at the federal level with the GPRA indicators.

Some examples of indicator statements follow. Parentheses contain elements that should be included, such as the specific subject under study (e.g., adults who have completed 100 hours of instruction), the measure (e.g., CASAS), the performance or behavior (e.g., increased score 3 points), etc.

- **For preschool-age children:** 60% (program expectation) of preschool children (subject) will increase a minimum of one percentile ranking on the auditory comprehension or expressive language subscales (behavior or performance) as measured by the Preschool Language Scales (PLS) (measure) after eight months of participation (time period).
- **For adult learners:** 60% (program expectation) of adult learners who participate in adult literacy services for at least 100 hours (subject) will demonstrate an increase of four or more points over their entry-level test scores (performance or behavior) on each CASAS subtest (measure).
- **For English Language Learners:** 50% (program expectation) of all parents who pre-tested at levels 1, 2, or 3 on the BEST and who have completed at least 100 hours of instruction for English Language Learners (subject) will demonstrate a one level gain (behavior or performance) as measured by the BEST (measure).

[1] Throughout the *Guide* we refer to the state-required indicators as "performance indicators." The statute calls them "indicators of program quality." Many states use different terms for required indicators, such as "performance standards," "state literacy indicators," "quality indicators," etc.

- **For school-age children:** 90% (program expectation) of third grade students in families that have reached enrollment status (subject) will be promoted to the next grade level (behavior or performance) as reported by the End of School Year Progress Report (measure).

All of the required Even Start indicators measure learning. The intent of the statute was for each state to develop a set of measurable learning outcomes that reflect high-quality family literacy practices. Indicators also often align with state standards for adult and K-12 education. For example, several states used the state attendance rate to set expectations for school attendance. The challenge for each state is to develop a set of indicators that fit its context.

In addition to the areas mandated by Congress, state coordinators may develop indicators for other outcomes. For example, states commonly add readiness indicators for 3-5 year olds, adult learning outcomes for English language learners, and parenting education outcomes. Where possible, these indicators are tied to existing state standards. For example, preschool indicators are usually tied to state-level early childhood standards.

Basic Ingredients of Indicators

Performance indicators describe individual achievements in program outcomes as a result of participation in the program. A complete indicator statement includes:

- Subject or audience
- Time period for measurement
- Performance or behavior expected of participant, including threshold level or criterion
- Measurement or assessment instrument
- Benchmark or expectation of program performance

Further examples of additional outcome indicator statements:

- **Parent-child interactive literacy activities:** 60% (expectation) of parents in families that have reached enrolled status (subject) will read to or look at books with their children three to four times a week (performance or behavior), as reported by pre-test and post-test parent-child literacy activities (measure).
- **Parents' support for children's learning:** Within in a program year, 80% (expectation) of parents who participated a minimum of 55 hours (subject) will visit the local library regularly to borrow books or other reading materials (performance or behavior) as measured by a parent-child family literacy rating scale (measure).
- **Adult English language learners:** 50% (expectation) of adults who have completed 75 hours of adult education instruction within the program year and pre-tested at Levels 0-3 on the English As A Second Language Oral Assessment (ESOLA) (subject) will demonstrate an increase of one grade level (performance or behavior) as measured by the ESOLA post-test (measure).

State indicators vary in their levels of detail and uniformity, especially in assessment instruments. For example, some state-developed indicators of literacy development for 3-5 year olds specify the instrument, others offer a choice of instruments, and still others provide no or vague guidance.

Role of project director. Project directors contribute their knowledge of the local context and provide access to local information. For example, if the evaluation requires access to school records, the project director will most likely have or establish contact with school staff.

Project staff members also have an overall sense of how instruments can be applied in their program. Project staff and evaluators usually decide on the instruments together.

Role of evaluator. Evaluators work with staff members to develop effective and efficient systems for collecting indicator data. The evaluator will likely spend a lot of time during a new project's first year setting up data collection procedures that are compatible with staff procedures. Over time, the effort of developing data collection procedures typically diminishes, and staff members will probably pay more attention to identifying local evaluation information that supplements and enriches the data required by the state.

Another important evaluator role is analyzing and interpreting state indicator and other project data so that results are meaningful and useful for project staff. Often this analysis involves disaggregating the data into meaningful subgroups, such as age, hours of participation, etc. (See page 30 for further discussion on using outcome indicator data to demonstrate program effectiveness and guide program improvement.)

Choosing Additional Outcomes at the Local Level

Local Even Start project directors and evaluators may also go beyond state-required indicators to measure project-specific outcomes. Given the variety of populations and contexts of local projects, general state-required performance indicators may not be appropriate for some groups served by the project, such as teen parents, migrant families, etc. Project directors may also want to add indicators to enhance the measurement of state-required learning outcomes. The project objectives listed in the application are a good source of additional participant outcomes to measure. For example, for school age children, outcomes beyond attendance and promotion could include grade changes in Math and English Language Arts.

Some examples of additional project-level education outcomes include:

- **For reading readiness:** Even Start kindergarten children will respond correctly to 50% or more of the items on the Concepts About Print test.
- **For reading:** 75% of children served by Even Start will continue to read at or above grade level in grades 3 and 4.
- **Gains in parenting literacy skills:** 60% of parents who score at Levels 1 (has knowledge) and 2 (able to understand) on the Family Literacy Parenting Scale will advance to the next level on the post-test.

In choosing outcomes it is important to consider:

- Is the outcome one the program can influence in a significant way?
- Will measuring the outcome help staff identify aspects of program implementation that can be changed?
- Do staff and other community stakeholders view the outcomes as important?

Role of project director. Staff members are instrumental in knowing and articulating the program aim. Although the local evaluator can facilitate staff discussion, the director and staff know the context and details of their program. The director also knows the kinds of outcomes that are most important to the local population. For example, if a project has a large number of English language learners, outcomes related to English language acquisition will be most meaningful to the program and the community.

Role of evaluator. Evaluators help program staff identify and develop outcomes that reflect the program's purpose. Through discussion, evaluators and staff determine measurable expectations of participants after they have engaged with the program over a specific time period, such as a year.

Selecting Instruments

As noted, some indicators specify the assessment instrument; in other cases, the project may choose among instruments. In this case, the evaluator can highlight each instrument's strengths and weaknesses and help staff members choose the best instrument, given the project's context (such as participant characteristics and instructional goals) for measuring an outcome. It is unlikely that any one instrument will be ideal.

When indicators call for performance data, such as a General Equivalency Diploma (GED) or school attendance rates, the evaluator may simply design a reporting mechanism rather than an instrument.

Evaluators need to consider several factors when selecting and using instruments to measure outcomes. Perhaps most important, evaluators should choose an instrument that permits individual outcomes to be aggregated, or combined, so that project staff and the evaluator can discuss the program as a whole. Another consideration is the appropriateness of the instrument for both the pre-test and post-test because a typical approach for measuring change is to look at differences between pre- and post-test scores.

Norm-referenced instruments. Norm-referenced instruments measure knowledge and skills in reference to a comparison group. This means that scores can be compared to a norm, or reference, group according to certain characteristics, such as age, grade-level peers, etc. Norm-referenced tests are helpful for measuring preschool outcomes because they provide a comparison for tracking learning development.

Examples of norm-referenced assessment instruments:

Children:

- *Peabody Picture Vocabulary Test 3rd Edition (PPVT-III)*
- *Preschool Language Scales (PLS-4)*
- *Test of Early Language Development 3rd Edition (TELD-3)*
- *Test of Early Reading Ability 3rd Edition (TERA-3)*

Adult:

- *Slosson Oral Reading Test*
- *Test of Adult Basic Education (TABE)*

Norm-referenced instruments are standardized so they are administered and scored consistently. Norm-referenced instruments allow raw scores to be converted into standard scores, as well as percentiles or normal curve equivalents, allowing for comparisons. (Other types of scores based on conversions of standard scores are grade-equivalents and developmental ages.) Norms provide a basis for interpreting performance in relation to defined population or groups.

In using a norm-referenced test, consider the following:

- Is the norm group appropriate and large enough for the individuals served by the project? For Even Start, comparison norm groups should reflect the age group(s) served by the project. The more defined the ages, the more useful the information about development will be, especially for young children. For example, the PLS-4 norms are based on approximately 15,000 children from birth through 6 years.
  The sample size of the norm groups must also be large enough to yield meaningful comparative data. An instrument may claim to use 2,100 children in the norming population, but the subgroup of children who are the same age as a project's sample may only contain 50 children. Information about the norming population and subgroups will be found in the test's technical manual.
- Are the correct scores being used and reported? Raw scores should not be reported. Most commercial instruments include scoring manuals that explain how to convert raw scores and interpret gain scores.

Criterion-referenced instruments. Criterion-referenced instruments are appropriate for measuring learning progress in specific content domains (e.g., reading, math). To measure outcomes for school-age children, schools typically use criterion-referenced tests in grades one through three. These instruments are best used to assess current knowledge at a point in time and to determine what the student needs to learn next. Criterion-referenced instruments measure mastery of skills or content against a specific set of performance standards. Criterion-referenced instruments can generate rich information about a student's progress and content mastery.

Criterion-referenced instruments compare student performance against pre-determined criteria, such as performance benchmarks, whereas norm-referenced instruments compare the student with other students.

In using criterion-referenced instruments, bear in mind that:

- The guidelines for scoring or understanding skill levels should be clear enough so that judging performance is consistent, fair, and accurate.

- In general, criterion-referenced instruments are more useful in measuring individual outcomes rather than group, or program, outcomes.

Validity and reliability. Validity and reliability refer to technical characteristics of assessment instruments; issues of validity and reliability arise with both norm-referenced and criterion-referenced instruments. An instrument has validity when it matches the construct it is intended to measure and predicts performance accurately. An instrument has reliability if it consistently measures the construct it was designed to measure.

Even if the state indicators specify the instrument, evaluators may want to know more about it. The technical manual will provide information about a test's validity and reliability. Below is a brief description of validity and reliability in relation to selecting an instrument to measure participant outcomes.

The **validity** of an instrument's score answers the question, "How do we know the instrument is really measuring what we want it to?" Measures are considered valid only if the instrument is used for the purpose for which it is designed. Validity factors include:

- Do project staff and relevant stakeholders consider the instrument valid? At a minimum, staff members must be able to look at the items on an assessment instrument and understand what is being measured. One criterion for validity is consistency with how the underlying concept is used in the field.
- Does the instrument permit generalization from one task to another? That is, can the instrument predict performance in a larger domain, such as reading or writing success?

The **reliability** of an instrument's score refers to the measure's consistency and repeatability. Scores are reliable to the extent they remain consistent despite fluctuations in the testing situation (i.e., time or day of week), or with different forms of the instrument, or different administrators.

Without reliability, it is difficult to know with any confidence what a student can do. One challenge in assessing young children is their rapid pace of development, making their behavior and performance inconsistent from day to day. Few instruments for assessing children have high reliability.

Another reliability issue concerns scoring consistency. Achieving consistency in scoring requires professional development and training scorers in applying criteria or procedures consistently. That is, to be reliable, one person scoring the instrument should get the same result as another scoring the same instrument.

Examples of criterion-referenced assessment instruments:

Children:

- *Get it. Got It. Go!*
- *Dynamic Indicators of Basic Early Literacy Skills (DIBELS)*
- *Phonological Awareness Literacy Screening (PALS)*

Adult:

- *Basic English Skills Test (BEST)*
- *Comprehensive Adult Student Assessment System (CASAS)*

Example: Using Data to Demonstrate Effectiveness

Indicator: 60% of adults who pre-tested at levels of 0-8.9 on the TABE and completed 100 hours of adult education instruction in reading will demonstrate one level gain as measured by the TABE post-test.

Number who completed 100 hours of adult education and took the TABE pre-test: 19

Number who demonstrated at least 1 grade level gain on the TABE post-test: 13

Calculation: 13/19 x 100 = 68%

Finding: Program met expectation.

Role of project director. The project director collaborates with the evaluator to select assessment instruments. In some cases, staff members may be more familiar than the evaluator with instruments in certain content areas, such as adult education or early childhood education. In that case, the project director will suggest appropriate instruments.

The final selection of instruments is a mutual decision between the project director and the evaluator. Both bring expertise to the decision. Project directors may have content expertise and they know their staff's abilities. Evaluators will add expertise in assessing instruments and matching them with intended purposes.

Role of evaluator. The evaluator may help staff review the assortment of instruments used to measure outcomes. Are they adequate for the project's purposes? Are additional instruments needed for specific subgroups? Is an additional instrument needed to augment an instrument's validity? Can some instruments serve multiple purposes?

The evaluator's primary responsibility is to help interpret scores, but he or she also takes the lead in comparing various instruments and training (or overseeing the training of) staff members in how to use the instrument, how to score it, and how to convert scores. For example, if instrument scores are to be converted to percentiles, the evaluator needs to make sure staff members know how to convert raw scores and use percentile rankings. If the scoring conversion is especially complex, the evaluator may take responsibility for converting scores.

Using Outcome Data for Continuous Improvement

Projects typically demonstrate a program's success or effectiveness in relation to a standard or benchmark. For example, to claim success in helping adults improve their reading skills, a project's data must answer the question, "compared to what?"

The most common comparison for Even Start projects is between the performance indicator and actual results. (See chart, page 31, for an example of reporting outcome indicator data.)

Other comparisons to demonstrate a project's success to broader audiences include comparing a project's outcome data with a test's norms or with aggregated state results.

What if the outcome result is less than expected? Ideally, the findings will point to areas where change is possible, and the evaluator can conduct additional analyses. Two ways to do this are by making further comparisons and disaggregating group scores.

Making further comparisons. In addition to the comparisons listed above, other useful comparisons are looking at project scores over time and at pre- and post-test results for individual participants.

One benefit of required indicators is that project staff can compare similar outcome data over time. For example, a project may not have met the required state indicator, but its results show improvement over time. Staff can use this feedback to decide if current strategies are working.

Example: Comparing Indicator Data for Two Years	*2001-2002*	*2002-2003*
Indicator	*Program Results*	*Program Results*
A1. 50% of the adults who have completed 100 hours of adult education instruction in reading and who pre-tested at Levels A-D on the READ or 0–8.9 on the TABE will demonstrate one grade level gain as measured by the READ or TABE post-tests.	*59%*	*47%*
C1. 40% of the adults participating in Even Start who have earning a high school diploma as their primary goal and who score 9.0 or above in reading and math on the TABE will earn a high school diploma or pass the GED during the program year.	*21%*	*36%*
Summary: Adults in group A1 achieved state indicator measures the first year, but did not in the second. Examining the reasons for this decline could be useful. Adults in group C1 did not meet state indicators either year, but the program showed progress from the first year to the second. It would be useful to learn what contributed to this increase in order to build on these factors.		

Another approach compares pre- and post-test scores. This approach is typically used to determine if a change from one time to another is statistically significant—i.e., the results are unlikely to have occurred randomly.

Pre- and post-tests compare participants with themselves. The pre-test score shows what a student knew prior to enrolling in your Even Start project; the post-test score shows what he or she knows after exposure to the project or curriculum. The shift in learning from the pre-test to post-test may be a result of a program's intervention, such as curriculum, instruction, etc.[2]

[2] There are many threats to validity associated with a pre-test and post-test design that may limit making statements of attribution. Because this Guide is not intended as a technical methods or statistics document, readers are encouraged to seek appropriate resources for a more thorough description of the limitations of pre-test and post-test analyses.

Taking sample size into account. Test results must be interpreted in light of the sample size. Typically, statistical tests are used with sample sizes of at least 30. For many Even Start programs, small sample size is a problem. Small sample sizes make it difficult to detect meaningful relationships and significant differences in learning. Findings from the statistical analysis may show no differences between groups when there really is a difference. One benefit of the pre-test and post-test design is that statistical t-tests are appropriate for small (fewer than 15) sample sizes. Another solution is to collect longitudinal data and study patterns over time. Another is to collapse groups in order to find patterns. For example, the results for several age groups such as 2 to 5 year olds can be combined.

Disaggregation by group scores. Another way to analyze outcome data is to sort data within the project population by meaningful sub-groups, such as participants who achieved the outcome and those who did not. These sub-groups may then be compared on relevant factors such as demographic characteristics (e.g., age, gender, race/ethnicity), contact hours, duration, entry level scores, opportunities for learning, etc.

Interpreting data. When interpreting data, the purpose of the analysis should be central. The evaluator and program staff will determine if the data are meaningful. The local evaluator arranges data so project staff can understand the findings and interpret the results. Chapter 11 presents approaches for reporting data. Below are issues evaluators should consider when interpreting data.

- Exercise caution in attributing results to causes. Although findings may show achievement gains, it may be hard to attribute them to the Even Start project. Other factors may have contributed to participants' achievements. Program staff and evaluators need to interpret findings with care and limit their conclusions to those the data can support. Using several data sources and methods can help rule out alternative explanations and demonstrate the project's influence.
- Statistical significance is not necessarily practical significance; statistically significant findings may have no real practical value. For example, findings may be statistically significant when there has only been a very small change in scores, such as ½ point. There may be a gain on a few test items, but the gains may not reflect meaningful progress.
- When making comparisons over time, the evaluator should compare the same participants, with the same intervention, over the same time period.

Armed with outcome data he or she feels confident about, the evaluator will likely next meet with staff members, present the findings, and guide discussions about what the findings mean for the project. Ideally, interpretations of outcome data should point to program areas that can be improved, such as professional development, curriculum integration, or the intensity of information. Some changes may be significant enough to become the focus of a new evaluation. Others may be less important, but all proposed changes should be documented so that their effects on participant outcomes can be tracked.

With the evaluator, the program leader should facilitate a staff discussion about understanding and using the data, and the project director should use the data consistently to make decisions, particularly about resources and program changes. Using data frequently, the project director can help staff members see the benefit of collecting data—an important step since staff will usually collect it.